CAICT 中国信通院　京东探索研究院 JD EXPLORE ACADEMY

No.202106

White Paper

on

Trustworthy Artificial Intelligence

China Academy of Information and Communications Technology

JD Explore Academy

July 2021

COPYRIGHT NOTICE

Foreword

As artificial intelligence (AI) technology continues to develop at a rapid pace, its influence has reached many aspects of people's lives and work. However, each coin has its other side. AI poses risks and challenges in addition to opportunities. "We must enhance the assessment and prevention of the potential risks posed by the development of AI to safeguard the people's interests and national security, making sure that AI technology is safe, reliable and governable," stressed by President Xi Jinping at the collective study session held by the Political Bureau of the 19th CPC Central Committee in October 2018. Enhancing confidence in AI applications and promoting the healthy development of the AI industry have become a top concern.

Developing trustworthy AI has become a global consensus. In June 2019, the Group of Twenty (G20) proposed the G20 AI Principles, emphasizing that the world should put people first during the development trustworthy AI. The principles have received widespread recognition from the international community. The EU and the US have also placed enhancing user trust and developing trustworthy AI at the heart of their AI ethics and governance efforts. An inevitable choice for the future world is to transform the abstract AI principles into concrete practices and apply them to technologies, products, and applications, thereby solving the society's concerns and preventing security risks. Developing AI trustworthiness is an important issue related to the long-term development of AI and an urgent task for the industry to prioritize on.

Whether reviewing the development background and history of trustworthy AI or looking into the future of new generations of AI, this White Paper believes that the stability, explainability, and fairness of AI systems will always be the

top concerns of all stakeholders. In view of the current situation, this White Paper begins from the perspective of implementing the global consensus on AI governance, focusing on trustworthy AI technology, enterprise and industrial practices, and analyzes the path to achieving the key trustworthy AI requirements of reliability & controllability, transparency & explainability, data protection, accountability, and diversity & inclusiveness. Meanwhile, this White Paper proposes recommendations on the future development of trustworthy AI.

As AI is still in a stage of rapid development, our understanding of trustworthy AI remains insufficient. Therefore, we welcome and appreciate the identification and correction of the shortcomings of this White Paper.

Contents

1. Development background of trustworthy AI 1

1.1. The risks of AI technology have resulted in a trust crisis 1

1.2. The global community has attached great importance to trustworthy AI 2

1.3. Trustworthy AI needs to be guided by systematic approaches 6

2. Trustworthy AI Framework 7

3. Support technologies of trustworthy AI 11

3.1. AI system stability technologies 11

3.2. AI explainability enhancement technologies 12

3.3. AI privacy protection technologies 13

3.4. AI fairness technologies 14

4. Practical approaches to trustworthy AI 15

4.1. Enterprise-level 15

4.2. Industry-level 21

5. Suggestions for trustworthy AI development 22

5.1. The process of AI regulation and legislation in China shall be accelerated at the Governmental level. 22

5.2. A comprehensive systematic, and forward-looking layout shall be made at the academic level 23

5.3. Enterprises' practices must adapt to the business development and achieve agile trustworthiness 23

5.4. A communication and cooperation platform shall be set up to create a trustworthy ecosystem at the industry organizations level 24

References 25

Figures

Figure 1 NUMBER OF RESEARCH PAPERS ON TRUSTWORTHY AI 3

Figure 2 TRUSTWORTHY AI PRACTICES OF THE ENTERPRISES 5

Figure 3 CORE CONTENT OF TRUSTWORTHY AI 7

Figure 4 GENERAL FRAMEWORK OF TRUSTWORTHY AI 8

Figure 5 KEYWORDS IN THE 84 AI ETHICS DOCUMENTS WORLDWIDE 10

Tables

Table 1 Common inherent biases in data sets 20

1. Development background of trustworthy AI

As a critical driver of revolutionary technological advancement and industrial transformation, AI empowers significant socio-economic transformation and profoundly influences the structure of the global political economy.

The global AI industry continues to grow steadily in 2020 and is valued at USD 156.5 billion, with an increase of 12% over the previous year, according to the IDC. Contributing to this is the value of the domestic Chinese industry at USD 43.4 billion (RMB 303.1 billion), with a year over year increase of 15% as estimated by CAICT.

Significant risks and challenges accompany the tremendous opportunities ushered in by AI. Addressing these issues, president Xi Jinping emphasizes the importance of "Safe, reliable and governable" AI and champions the adoption of G20 AI guidelines to facilitate robust AI development globally.

1.1. The risks of AI technology have resulted in a trust crisis

The increasing length and breadth of AI utilization is making it a key component of the information infrastructure. However, latent risks continue to emerge in the process, which are reflected in the following aspects:

Application risks caused by algorithm security. Due to the vulnerability of deep learning-based AI technology to interferences and targeted attacks, it is difficult for the public to have confidence in the reliability of AI-based systems. Negative examples include the death of pedestrians caused by Uber's autonomous cars, which failed to identify them on time and an AI company's successful bypass of many countries' facial recognition systems through spoofing attacks with 3D masks and composite photos[1].

The black box model results in algorithmic opacity. The complexity and unpredictable outcome of deep learning algorithms introduce risks associated with uncertainty. Additionally, the lack of an intuitive understanding of an AI's decisions has prevented the greater integration of AI with traditional industries. For example, a school in Texas in the United States used an AI-based system to evaluate teaching performance. However, the school had to take the system offline following complaints by teachers over its incapacity of interpreting the judgment basis for controversial decisions.

Data discrimination leads to intelligent decision-making biases. The trained AI model is prone to input bias. Therefore, if biased data is used for model training, the algorithm may amplify the bias, subsequently influencing the AI's decisions. For example, the Correctional Offender Management Profiling for

[1] https://new.qq.com/omn/20191230/20191230A0FX0R00.html

Alternative Sanctions (COMPAS) system used by Chicago courts in the United States has turned out to be discriminatory towards black people[2].

Complex system decision-making makes it difficult to define the subjects of liabilities. The autonomous AI-based decision-making process is influenced by numerous factors, making it difficult to determine accountability. Given the frequent incidents caused by autonomous driving, robots, and related applications, law experts have suggested that although the AI is unlikely to be held accountable for its actions based on current laws, further discussion is still needed to decide whether the owner of the AI or its software developer should be held accountable in the case of an infringement[3].

Data abuse leads to privacy leakage risks. The frequent use of biometric authentication increases the risk associated with potential data breaches, which could cause the leaking of confidential user data.

For example, ZAO's illegal collection of facial data via its user agreement[4] has raised concerns over the misuse of private data, which may cause increased risks associated with identity theft which can compromise biometric authentication and financial services.

1.2. The global community has attached great importance to trustworthy AI

In the face of a potential crisis, the global community has reached a consensus on developing trustworthy AI. In June 2019, in its G20 AI Principles, the Group of Twenty (G20) proposed five recommendations for Governments to follow. Most notable of which include enhancing public and private investment in AI in the hopes of promoting innovation in Trustworthy AI and creating a strategic policy-making environment that will encourage the development of trustworthy AI. These have become widely accepted as the guideline for AI development by the international community.

The academic community has pioneered explorations in trustworthy AI. In November 2017, Chinese scientist He Jifeng, a fellow of the Chinese Academy of Sciences, first introduced the concept of trustworthy AI in China at the S36 Session Seminar of the Xiangshan Scientific Conference, pointing out the inherent trustworthy qualities of AI.

From the perspective of academic research, the research scope of trustworthy AI includes stability, explainability, fairness and privacy protection. The number of published research papers on trustworthy AI in

[2] https://www.sohu.com/a/299700146_358040

[3] http://media.people.com.cn/n1/2018/0502/c40606-29959959.html

[4] http://finance.china.com.cn/industry/company/20190909/5075700.shtml

2020 increased by nearly fivefold over 2017. Defense Advanced Research Projects Agency (DARPA) published an academic report titled Explainable Artificial Intelligence and initiated funding activities to promote the development of trustworthy AI. In addition, a top-level conference on AI, the AAAI has organized a symposium on Explainable AI for two consecutive years and has maintained the active research trend. Furthermore, the FAccT ML

（Fairness, Accountability and Transparency in Machine Learning）

community was formed to explore the fairness, accountability and transparency of machine learning. Since 2018, the ACM has held the ACM FAccT (ACM Conference on Fairness, Accountability, and Transparency) Seminar, which has been ongoing for four consecutive years.

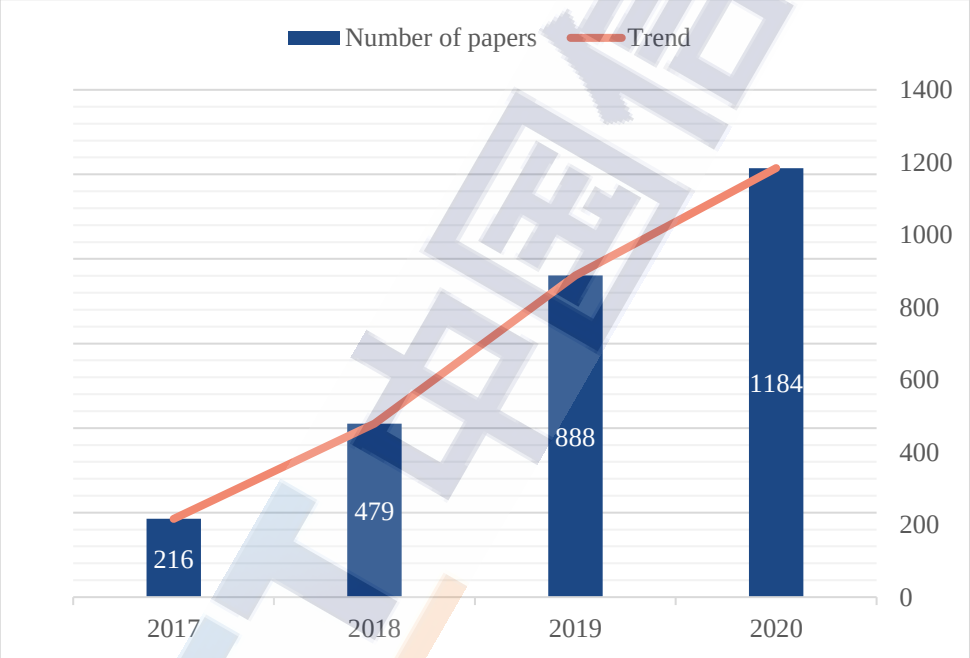

FIGURE 1 NUMBER OF RESEARCH PAPERS ON TRUSTWORTHY AI[5]

Source: Web of Science

Governments have placed enhancing user trust and developing trustworthy AI at the heart of their AI ethics and governance efforts. In 2020, the EU proposed in its **White Paper on Artificial Intelligence [1], the trustworthy AI ecosystem, aiming to implement a European AI regulatory framework and propose mandatory regulatory requirements for high-risk AI systems.** In December of the same year, the White House published an administrative

[5] Retrieved and sorted by CAICT according to Web of Science.

order[6] titled Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government to set guidelines for the use of AI by federal agencies, aiming to promote public acceptance and trust in the Government's use of AI technology in decision-making.

Standardization authorities have formulated trustworthy AI standards. Internationally, ISO/IEC JTC1 SC42 has established the Trustworthy AI Working Group (WG 3), which released the Information Technology – Artificial Intelligence – Overview of Trustworthiness in Artificial Intelligence standard. The standard is helping to advance progress on a series of research areas ranging from Information Technology, Artificial Intelligence to evaluating the robustness of Neural Networks. Similarly in China, the Artificial Intelligence Subcommittee of the National Information Technology Standardization Technical Committee (SAC/TC 28/SC 42) was established to promote related research. In November 2020, the National Information Security Standardization Technical Committee TC 260 WG issued the Practice Guide to Cybersecurity Standards – Guidelines on the Code of Ethics for Artificial Intelligence (Draft for Comment). In view of the possible ethical problems of artificial intelligence, the Draft proposed the norms and guidelines for carrying out AI-related activities safely.

Enterprises have actively explored and implemented trustworthy AI. IBM Research AI developed several AI trust toolkits in 2018 to evaluate and test the fairness, robustness, explainability, accountability, and value consistency of AI products in the development process. These toolkits have been donated to the Linux Foundation as open-source projects. Similarly, other Chinese and foreign enterprises, including Microsoft, Google, JD.com, Tencent, and Megvii, also actively engage in related activities. Figure 2 summarizes the exploration in the field of trustworthy AI by some enterprises.

[6] https://www.thepaper.cn/newsDetail_forward_10263830

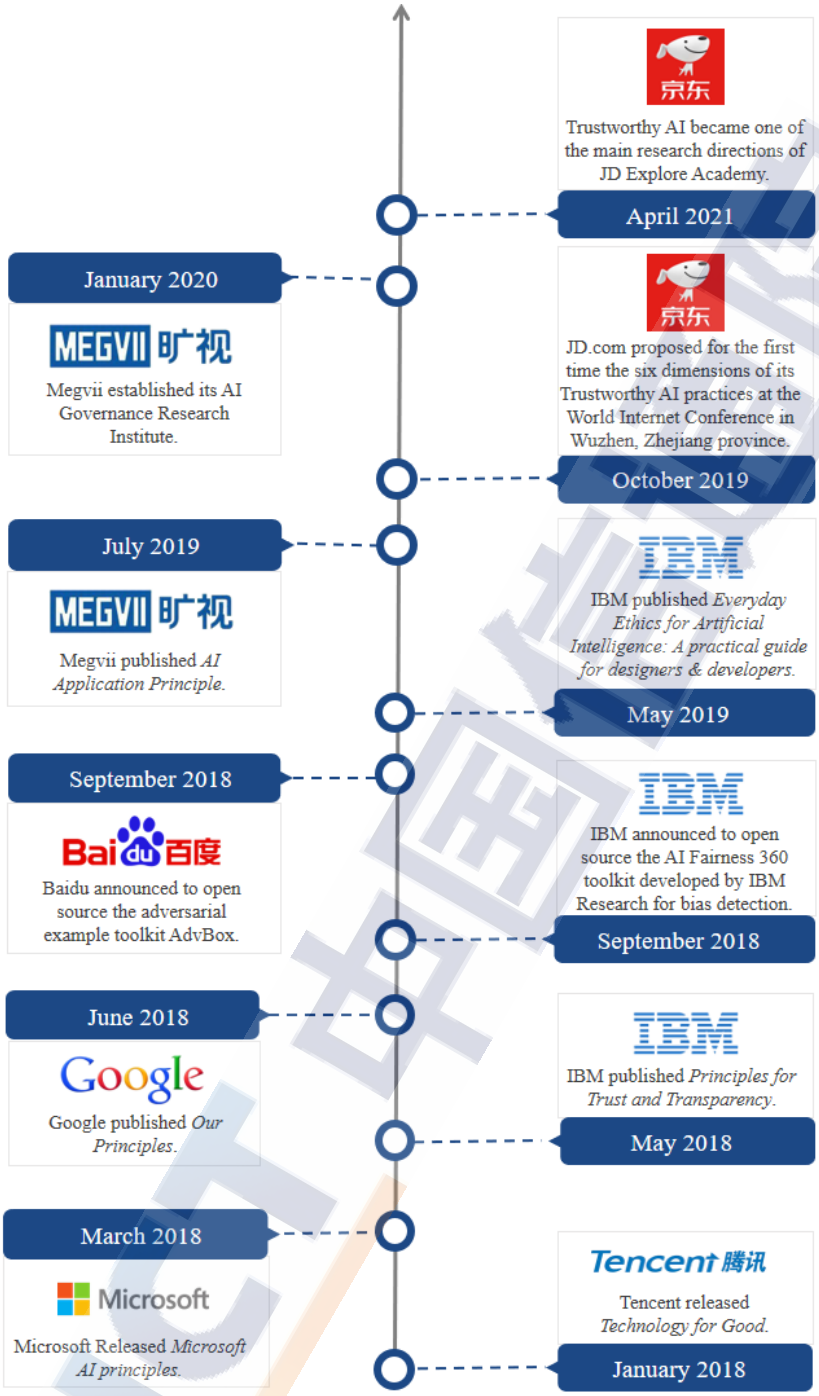

FIGURE 2 TRUSTWORTHY AI PRACTICES OF THE ENTERPRISES[7]

Source: Data compilation

Summarizing the descriptions from multiple sources, this White Paper believes that "trustworthiness" reflects the credibility of a series of internal attributes of AI systems, products and services, such as safety, reliability, explainability and accountability. **In terms of technology and engineering practices, trustworthy AI can implement ethical governance requirements and**

[7] Organized from disclosure information

effectively strike a balance between innovation-oriented development and risk governance. With the accelerated development of AI-based technology and industry, the implications of trustworthy AI will continue to be expanded upon in the future.

1.3. Trustworthy AI needs to be guided by systematic approaches

The requirements for trustworthy AI and the practicality of assessment methods have been increasingly strengthened. All countries have noticed that, ethics washing is likely to occur if there is no corresponding enforcement mechanism for "soft" ethical constraints.[8] Therefore, more practicable policies are needed. In February 2021, the German Government issued the AI Cloud Service Compliance Criteria Catalogue (AIC4)[9] and defined the trustworthiness assessment criteria for AI in a cloud environment from a practical level. In April 2021, the European Commission announced its proposal for a regulation laying down harmonized rules on artificial intelligence and amending certain union legislative acts (Artificial Intelligence Act). The proposal adopts a four-level risk-based approach whereby the uses of artificial intelligence are categorized according to human safety and fundamental rights, and prescribing corresponding penalties. The proposal aims to build market trust by enhancing the legal framework, thereby promoting the widespread adoption of AI technology and boosting confidence in AI. In May, the National Institute of Standards and Technology (NIST) proposed a method to evaluate user trust in AI systems, publishing the Artificial Intelligence and User Trust (NISTIR 8332) guideline[10], which outlined the practical assessment criteria for trustworthy user experiences when interacting with AI systems. In June, the U.S. Department of Defense strengthened its commitment to building reliable AI capabilities through education and training, and adopted systems engineering and risk management methods to implement supervision throughout the procurement life cycle.

Despite the development of legislative frameworks for AI, detailed rules and regulations still need to be specified. Meanwhile, the industry has entered a deep water zone where tough challenges of the practice must be met. Generally, **the practices of trustworthy AI remain scattered. A systematic methodology is needed** to fully implement the relevant governance requirements and guide the implementation of relevant operations. Based on a comprehensive review of the ethical constraints, laws, regulatory norms, and good practices on AI, this White Paper **proposes a Trustworthy AI Framework** that outlines methodologies for implementing AI governance

[8] https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/

[9] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html

[10] https://www.nist.gov/news-events/news/2021/05/nist-proposes-method-evaluating-user-trust-artificial-intelligence-systems

requirements from the industry's perspective. The proposal is based on an in-depth analysis of trustworthy practices in use by enterprises and industries and aims to bridge the gap between AI Governance and industrial practice.

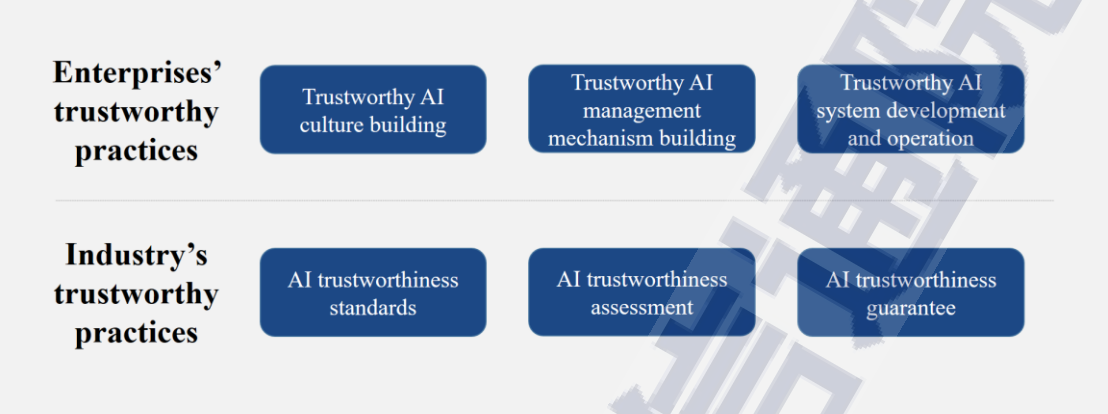

FIGURE 3 CORE CONTENT OF TRUSTWORTHY AI

Source: CAICT

In terms of trustworthy enterprise practices, the framework proposes practical requirements for each stage of the AI life cycle in conjunction with five essential trustworthy requirements. It also makes detailed recommendations on establishing a company culture and management mechanisms that help develop trustworthy AI. **In terms of the industry's trustworthy practices,** the framework addresses them from three perspectives: standards, evaluation, and safeguards.

2. Trustworthy AI Framework

From its initial proposal by academia to its exploration by various groups and its eventual industrial implementation, trustworthy AI's connotation has gradually evolved and become enriched. This White Paper argues that trustworthy AI can no longer be defined solely through the state of related technology, products, and services. Instead, its definition has expanded to include a systematic methodology that encompasses the creation of "trustworthiness" from all aspects. Figure 4 shows the general framework of trustworthy AI.

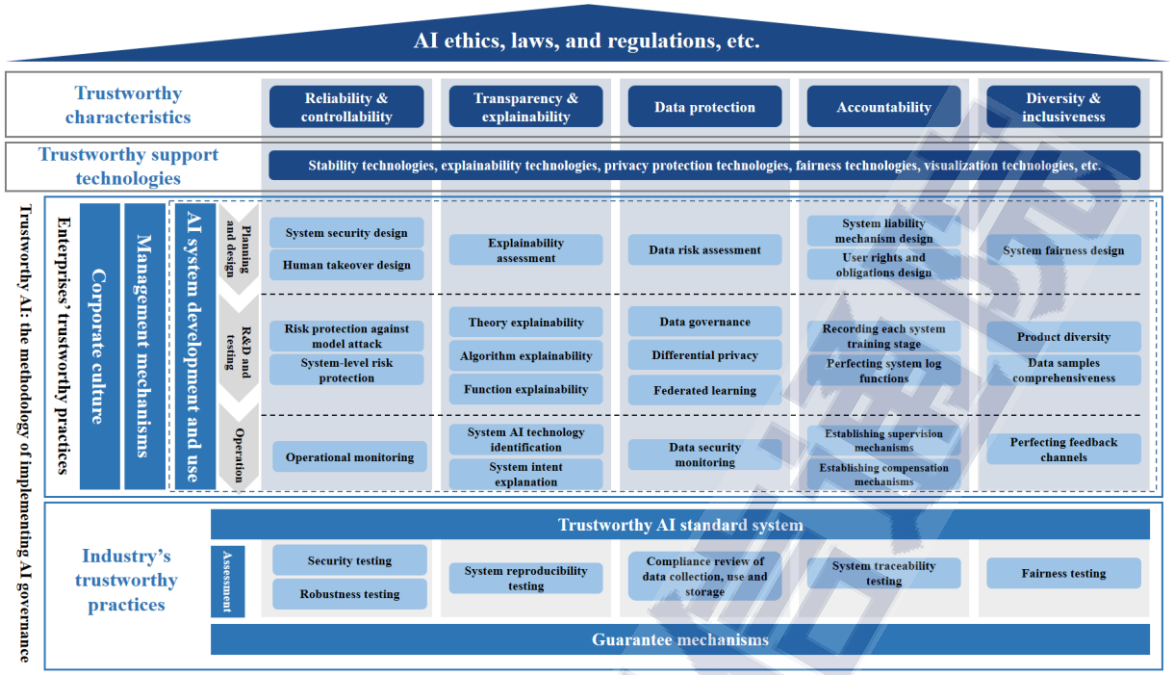

FIGURE 4 GENERAL FRAMEWORK OF TRUSTWORTHY AI

Source: CAICT

Trustworthy AI is a crucial factor for the implementation of AI governance principles in practice. Its trustworthy characteristics align with the requirements of AI ethics, related laws and regulations, all of which are based on a human-centric approach.

From the perspective of governance, just as ethics provide guidance at the macro level, laws impose constraints based on specific results. Trustworthy AI penetrates from all aspects of an enterprise's operating and management activities to the relevant industry activities, in this way **transforming relevant abstract requirements into specific functional requirements applicable in practice**, so as to improve social trust in AI.

Trustworthy characteristics of AI. Based on a word frequency analysis of 84 international policy documents on AI governance, the trustworthy characteristics of AI are summarized by five main aspects: **transparency, security, fairness, accountability, and privacy** [2]. Due to diverse cultural backgrounds, varying business nature, and management systems, each organization may have a unique understanding and different realizations of these common principles. However, from the industry's overall perspective, **the aforementioned five core concepts have been proposed and refined based on a consensus over the ideal methodology to build a multilateral standard for trustworthy AI**. The five consensuses provide guidelines on

strengthening trust in the use of AI on both the supply and demand sides and how to assist regulators in fostering a trustworthy and healthy industrial ecosystem. Referencing the five consensuses, the Joint Pledge on Artificial Intelligence Industry Self-Discipline [3] and the Trustworthy AI Operational Guideline [4] initiated and issued by the China Artificial Intelligence Industry Alliance (AIIA), this White Paper has summarized and put forward five elements: **reliability & controllability, transparency & explainability, data protection, accountability, and diversity & inclusiveness.** These elements help outline the operational capacity required for the practical application of trustworthy AI.

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, Excitability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety,harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability,liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being,peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, autonomy choice, self-determination, liberty,empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment(nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

FIGURE 5 KEYWORDS IN THE 84 AI ETHICS DOCUMENTS WORLDWIDE[11]

Source: Data compilation

In terms of the trustworthy support technologies, the inadequacies of current technologies shall be solved through theoretical research and technical innovations underpinned by the five trustworthy characteristic requirements: reliability & controllability, transparency & explainability, data protection, accountability, and diversity & inclusiveness. Examples include researching a new generation of explainable AI algorithms and investigating computing

[11] Source: *The Global Landscape of AI Ethics Guidelines*, sorted by CAICT.

technologies capable of privacy protection. Such exploration requires active and collaborative efforts from the academic and industrial communities.

In terms of the enterprises' trustworthy practices, their implementation of trustworthy AI is crucial to the development of methodologies for building trustworthiness. Since the widespread adoption of AI technology is dependent upon commercialization by enterprises and other market entities. Despite the flaws inherent to many technologies, the ability to leverage their potential while avoiding misuse is critical to success. To ensure success, the enterprise must cultivate a culture built around trustworthiness and implement trustworthy management systems. Furthermore, trustworthy features need to be implemented throughout the lifecycle of AI system development and deployment, ensuring that trustworthy qualities are innately built into a product since its conception.

In terms of the industry's trustworthy practices: Building trustworthy AI requires the entire industry's participation through practical implementation. Methodologies include the establishment of trustworthy AI standards, the evaluation of AI trustworthiness, and the implementation of trustworthy AI safeguards. Furthermore, the industry can use insurance and other social mechanisms to share the risks associated with AI technology and system deployment.

3. Support technologies of trustworthy AI

With the increasing public focus on the trust problems of AI, safe and trustworthy AI has become the trending topic of research [5,6]. The research is focused on promoting the stability, explainability, privacy protection, and fairness of AI systems. These technologies establish the foundation of trustworthy AI.

3.1. AI system stability technologies

AI systems and their data face series of unique interferences, including poisoning attacks, adversarial attacks, and backdoor attacks. These attacks can occur individually or in unison. For example, poisoning attacks can inject interference data into the training data set through malicious comments to degrade the accuracy of the recommendation system [17]. Adversarial attacks can attach adversarial patterns on traffic signs, causing the automatic driving system to misidentify the information on the signs, thus causing traffic accidents [18]. Finally, backdoor attacks are hidden and may be used to attack the AI supply chain. Compared with conventional software systems, AI systems must have a higher tolerance to such interference.

The stability of AI system has attracted sustained and extensive research. Adversarial and poisoning attacks against AI models have emerged as early as 2012 and 2013. The purpose of adversarial attacks is to induce decision-making

errors in AI systems by specifically designed targeted sample data. The purpose of poisoning attacks is to degrade the performance of the trained model by injecting malicious data into the training data set of the AI model. Following the conception, adversarial attacks have successively evolved into the fast gradient sign method (FGSM), Carlini-Wagner method and projected gradient descent method (PGD). Poisoning attacks also developed rapidly, giving rise to backdoor attacks. A backdoor attack injects a backdoor into the AI system through backdoor samples, enabling hijacking of the system. Backdoor attacks have some similarities with poisoning attacks and often inject backdoors into the system through poisoning attacks. To resist these attacks, various abnormal data detection methods have been proposed to detect and remove malicious data, including adversarial samples, poisoning samples and backdoor samples, thus mitigating the interference caused by malicious attacks. Besides, adversarial attacks can also be resisted via adversarial training. Finally, backdoor attacks can be resisted through model pruning and backdoor detection.

Nonetheless, numerous challenges lie ahead for AI stability. On the one hand, various new interference techniques have emerged and continue to evolve, while new attack methods may quickly nullify the old defense techniques. On the other hand, the interference forms are gradually spreading from the digital world to the physical world. For example, the attacks can affect autonomous driving and facial recognition systems directly on the physical level through printed adversarial examples. With this in mind, we predict that research on AI stability will continue to increase in the future.

3.2. AI explainability enhancement technologies

Currently, AI systems using deep learning-based algorithms operate like black boxes. Observations can only be made of the data input and output but the internal working principles and judgment basis remain unclear. On the one hand, it is unclear why the trained AI models can have such a high performance. On the other hand, the factors that AI systems rely on to make decisions are unclear.

Research on the explainability of AI algorithms is still in its infancy, and the theoretical framework of some algorithms need further refinement [7,8,9,14,15]. For example, the effectiveness of optimization algorithms on simple AI models such as decision trees and support vector machines has been proved. However, despite the extensive body of research on the high efficacy of stochastic gradient descent algorithms at optimizing deep neural networks, the discussion is ongoing, and this issue remains inconclusive. In another example, research on how AI models utilize data features to make predictions have yielded promising results through experiments conducted by academia. However, the theoretical foundation is yet to be clearly defined. To improve the explainability of AI models, researchers have suggested: designing suitable

visualization techniques to aid in the assessment and explanation of intermediate model states; employing influence functions to deduce the influence of training data on the final convergent AI model; using the Grad-CAM (Gradient-weighted Class Activation Mapping) method to analyze the data features used by AI models when making predictions; adopting the LIME (Local Interpretable Model-agnostic Explanations) method to use simple interpretable models to approximate complex black box models locally and further exploring their explainability; improving model reproducibility by perfecting management mechanisms for model training.

In industrial applications, AI system behavior should be as transparent, explainable, and trustworthy as possible to human beings. Trusting decisions made by AI systems without a clear explanation of the decision-making process will significantly limit the widespread adoption of these systems in key fields, such as national defense, law, medical care, and education, and potentially cause serious social problems. Therefore, enhancing the explainability of AI systems is a matter of great urgency.

3.3. AI privacy protection technologies

Artificial intelligence systems rely on massive amounts of data to make decisions. However, the data transfer process and the AI models themselves are vulnerable to the leakage of sensitive and private data. For example, at any stage of the data transfer process, attackers can attack the anonymous data sets to steal data. In the data publishing stage, attackers may use re-identification attacks to attack anonymous data sets and steal private information. Attackers can also attack the AI model directly and steal private information. For example, model inversion attacks can infer and reconstruct the training data according to the output of the targeted model to steal private information; membership inference attacks can infer whether a given data sample is from the training data set of the targeted model, which thus causing sensitive data to be compromised.

Scholars have proposed various tailored protection methods for the above privacy leakage problems, with methods based on differential privacy and federated learning being most common. Differential privacy is first proposed by the American scholar Cynthia Dwork [10] in 2006. It is a major quantitative indicator of the privacy protection capability of AI systems. The differential privacy follows the idea that an AI algorithm with excellent privacy protection capability should be insensitive to small perturbations in the input data. Based on this idea, privacy attacks can be prevented through techniques including down sampling, permuting, and adding noise. In a work proposed by Google in 2016, differential privacy was applied to improve the privacy protection capability of deep learning models for the first time via adding Gaussian noise to the gradients during model training. This work demonstrated the potential of applying differential privacy methods to large-scale AI models. At present,

some leading technology companies have applied the differential privacy method to parts of actual business operations. Proposed in 2015, federated learning [19] aims to train AI models without collecting user data to protect private information. Specifically, federated learning first deploys AI models onto user devices whereby each user device uses its own private data to calculate the gradient of model parameters and upload it to the central server. Second, the central server integrates the collected gradients and sends them back to each user device. Finally, each user device uses the integrated gradient to update the model. However, it should be noted that the federated learning approach is still subjected to the risk of private data leakage, according to some preliminary studies. Experiments have shown that federated learning could leak a small amount of local user data [11] while some theories state that federated learning may weaken the privacy protection capabilities of AI systems to some extent [12]. Therefore, federated learning needs to be further optimized to improve its user privacy protection capability. A feasible direction is to combine federated learning with differential privacy to build an AI system with stronger privacy protection capabilities.

In the current era, data carries more and more private information. Users have begun to pay more attention to the protection of private data than ever before. Some countries have also made legislative efforts to formulate usage specifications for private data. Directing research efforts on privacy protection can allow AI systems to conform to the basic rules and regulations stipulated by law, thereby complementing the establishment of trustworthy AI.

3.4. AI fairness technologies

Following the widespread usage of AI systems, problems including unfair decision-making and discrimination against certain groups have emerged. In academia, it is argued that the main reasons behind the occurrence of decision-making bias are as follows: The distributions of data in the data sets are unbalanced due to limited data collection conditions; Training on unbalanced data sets may cause AI models to sacrifice the performance on a small amount of data for the sake of performance improvement on the overall data, resulting in unfair decision-making by the trained models.

To ensure the fairness of AI systems in decision-making, researchers in the field have made efforts through the following methods: minimizing the inherent discrimination and bias in the data by building a complete and heterogeneous data set; checking the data set periodically to maintain high data quality; using algorithms based on quantitative indicators of fairness to reduce or eliminate decision-making deviation and potential discrimination. The existing fairness indicators can be divided into two categories, individual fairness and group fairness [13,16,20]. Individual fairness and group fairness measure the degree of bias intelligent decision-making has towards different individuals and different groups respectively. Furthermore, algorithms based on

fairness indicators can be approximately divided into three categories: pre-processing methods, in-processing methods and post-processing methods. Pre-processing methods clean the data by sensitive information removing, resampling, and other strategies, thereby reducing the deviation in the data. In-processing methods improve model fairness by adopting regularization terms that can quantitatively represent fairness during the training process of the AI model. For example, there is work proposed to use Rényi correlation as a regularization and further reduce any potential correlation between model prediction and sensitive attributes via min-max optimization. Post-processing methods can improve a trained model's fairness by adjusting its output. For example, based on the concept of multiaccuracy, the Multiaccuracy Boost method was proposed to reduce the decision deviation in black box AI systems.

As the utilization of AI in sensitive areas including recruitment, criminal justice, and medical care continues to grow, the fairness of AI has raised widespread concern. Fairness technologies can balance the data technically to guide AI models to deliver fair results, which will help improve the fairness of decision-making in AI systems.

Currently, an increasing number of studies are focused on the challenges in stability, explainability, privacy protection, and fairness of AI. As the in-depth studies continue, they will contribute to more stable, more transparent, and fairer AI theories and technologies which will become the cornerstone and important guarantee for the realization of trustworthy AI.

4. Practical approaches to trustworthy AI

By referring to the *Trustworthy AI Operational Guideline* issued by the Artificial Intelligence Industry Alliance (AIIA) in combination with insight gained on the state of research and development in AI enterprises through interviews, this White Paper has summarized and proposed a path towards the realization of trustworthy AI from the enterprise and industry level.

4.1. Enterprise-level

Enterprises are the core subjects of researching, developing and using the AI technologies, products and services, as well as the most important participants in the implementation of trustworthy AI practices. As an integral, evolutionary, and non-traditional system engineering project, the implementation of trustworthy AI practices in enterprises needs to start with adjusting corporate culture and management systems and fully implementing the relevant technical requirements during the development process.

4.1.1. Integrating trustworthy AI into the corporate culture

Corporate culture is the embodiment of an enterprise's overall values, shared vision, mission, and way of thinking. To develop trustworthy AI, an enterprise must integrate the concept of trustworthiness into its corporate culture.

(1) Management must endorse the "trustworthy" direction
As the core of an enterprise's operation, the management must agree on developing trustworthy AI, fully establish human-centric values and recognize the characteristic elements of AI trustworthiness: transparency & explainability, diversity & inclusiveness, reliability & controllability, accountability, and data protection. Furthermore, they must integrate trustworthy AI into all aspects of the enterprise's operation and management to promote the advancement of overall trustworthiness.

(2) Employees must reinforce "trustworthy" learning and practices
Enterprises can develop trustworthy AI-related learning and training programs by inviting experts as guest speakers and distributing books or introductory materials on trustworthy AI to disseminate the concept of “trustworthiness” among its employees. They must also promote the use of trustworthy technologies or tools and encourage employees to innovate and realize trustworthy AI in their work.

(3) Enterprises must create a "trustworthy" cultural atmosphere
Enterprises may reflect the elements of trustworthy AI in their offices, on websites, in publicity materials, and in news articles to exhibit their exploration of trustworthy AI in practice. This can motivate employees to discuss trustworthy AI topics and provide encouragement to the teams or individuals who have made contributions in the field.

4.1.2. Perfecting the management systems of trustworthy AI

Management systems can provide a basis for management actions and a guarantee for the smooth proceeding of social reproduction. Therefore, enterprises must perfect their management systems in a bid to realize trustworthy AI.

(1) Building trustworthy AI teams
Enterprises must build dedicated teams (or virtual organizations) to take charge of managing trustworthy AI. Such teams or organizations should be headed directly by the main personnel in charge to better command and coordinate other departments to participate in trustworthy AI-related work. The teams may be divided into sub-groups depending on the nature of the businesses. The members should be full-time or part-time personnel with legal and R & D backgrounds. The responsibilities of related departments and personnel must be clearly defined.

(2) Building and implementing trustworthy AI personnel management systems
The trustworthy AI management department must take the lead with the aid of other departments, including human resources, R & D and legal affairs to establish a management system for personnel working with trustworthy AI that should clearly outline the requirements on personnel management, education, training, and assessment. The targeted personnel are those involved in AI demand analysis, product design, research, development, testing, and

trustworthiness management. The enterprise must implement the personnel management system effectively and conduct education, training, and assessment regularly to gradually improve personnel expertise.

(3) Establishing and implementing management systems for the development and use of trustworthy AI systems

Enterprises must establish management systems for the research and development of AI systems and clarify the responsible departments, personnel, work content, work methods, work processes, and work requirements. The trustworthy AI management department will be tasked with overseeing the detailed implementation. Enterprises must also specify management systems for the use of trustworthy AI systems, and formulate emergency plans and relief measures to ensure that the systems meet the trustworthiness requirements during regular use and ensure that the damages and losses can be minimized or effectively solved in time when issues arise.

(4) Allocating necessary resources for trustworthy AI

Enterprises must coordinate and allocate the necessary resources for realizing trustworthy AI, including but not limited to personnel, funds, sites, and facilities.

(5) Establishing iteration and update mechanisms for the management systems

Given the constant changes in AI governance and the introduction of new policies and regulations, the enterprises' trustworthy AI management departments must take the lead in constantly optimizing and improving the management systems to adapt to the latest requirements to achieve the best results.

4.1.3. Embed trustworthy AI requirements into the whole process of R & D and application

(1) Planning and design stage

Enterprises must fully consider and implement the characteristic elements of trustworthy AI from the beginning of the AI system life cycle, firmly rooting the trustworthy concept in the critical stages of planning and design, including demand analysis and system specification design. Such efforts will ensure that the subsequent development, testing, and operation of the systems will continually meet the core requirements of trustworthy AI.

In conjunction with the established practices in software design, enterprises can build dedicated trust supervision teams to assist the production teams in developing trustworthy design schemes of AI systems from two aspects:

Putting forward the trustworthy design requirements for AI systems. Following product demand analysis, the trust supervision teams must comprehensively investigate the potential risks faced by the AI's intelligent systems and propose targeted countermeasures concerning system security, fail-sale mechanisms, explainability, data risks, system liability mechanisms,

user rights, user obligations, and system fairness. A list of trustworthy design requirements must be produced in the process.

Reviewing the trustworthy design schemes of AI systems. Specialists in the trust supervision teams must use their professional knowledge, work experience and prior case-study experience to verify the feasibility of the trustworthy design schemes to identify potential problems, provide insight, and propose further ideas on trustworthy design. Thusly, the experts can provide invaluable advice on the subsequent modification and improvement of the trustworthy design schemes to integrate core trustworthy AI characteristics with system design, reducing trust loopholes and prevent potential incident risk.

(2) R & D and testing stage

In terms of reliability & controllability, enterprises must strive to improve AI systems' defense capabilities and ensure human supervision and takeover powers. The defensive capabilities of AI systems may be improved from two levels: data and model. Data-level defense methods include bad data pre-cleaning and model robustness improvement via data augmentation. On the model level, aside from conventional defense methods that include model encryption and malicious query interaction limiting in the production environment, another important technique is adversarial training. Considering that AI models are vulnerable to tailored attack samples, adversarial training algorithms use adversarial examples to train the AI models, improve their robustness, and render them less susceptible to interference from the adversarial samples. Additionally, enterprises must establish backup plans for AI systems in the development process to ensure that they can be contained by failsafe measures when incidents occur following online deployment. AI system failsafe measures include automatic regulation and recovery, rapid access human overrides and a "one-key shutdown" manual procedure.

In terms of transparency and explainability, the key is to improve the reproducibility of AI systems. Currently, research on the explainability of algorithms is outpaced by the rapid development of AI applications. Therefore, enterprises must strive to improve AI system reproducibility in the R & D and testing stage. Improved reproducibility can enhance system transparency, reduce the difficulty of subsequent system audits and improve fault traceability. Relevant measures include: establishing a perfect data set management mechanism to record in detail the sources and composition of the training and testing sets used in the training process of each version of the system, as well as the data preprocessing operations adopted in the training process; establishing a perfect model training management mechanism to record in detail the hardware platform, system configuration, software framework, model version, model initialization, hyperparameters, optimization algorithms, distributed operation strategies, network rate, indexes, test results, and other techniques and engineering technologies used in model training.

In terms of data protection, enterprises must avoid problems, such as illegal collection, abuse, and leakage of training data, through the implementation of data governance measures and the exploration of privacy protection algorithms to train AI systems. Either differential privacy or federated learning technologies can be used to improve the privacy protection capabilities of AI systems from the algorithm level. Their application cases include Apple's user data collection and the U.S. Census. OpenDP, an artificial intelligence project co-founded by Microsoft and Harvard University, has developed many open-source differential privacy toolkits to protect models and data better.

In terms of accountability, enterprises must audit the implementation processes of AI systems comprehensively to improve their traceability and ensure the trustworthiness of the systems and services from the source. The main audit stages include data preparation, model training, and model assessment. An audit of the data preparation process can help to identify and confirm legal compliance in training data collection, the presence of privacy breaches, adherence to the standard annotation and preprocessing means when processing data, and the use of encryption, access restriction, and other security measures when storing data. Model training is the key to endowing AI systems with "intelligence". A comprehensive audit of the main training components, such as hardware platform, software framework, algorithm selection, and parameter tuning, can help to trace the systems. Model assessment can reflect AI systems' performance and generalization capabilities in practical applications to a large extent. A standard and rigorous assessment process can detect errors, measure model quality, and determine if design requirements are met, helping to track problems in the system implementation process for continuous improvement. Therefore, a detailed audit of the model's performance and parameter changes on the validation and testing sets is required.

In terms of diversity and inclusiveness, enterprises must focus on the fairness and diversity of the training datasets to avoid the lack of trust due to data biases. The performance of an AI system depends on the quality of the training data. The data sets may contain implicit racial, gender, or ideological biases (Table 1), leading the AI system to make inaccurate or biased, and discriminatory decisions. Enterprises must **improve the diversity and fairness of the training data** to meet the diversity and inclusiveness requirements. On the one hand, they must note the inherent discriminations and biases that may appear in the data and take proactive measures to reduce their impact. On the other hand, they must check data sets periodically to ensure high data quality. Additionally, **AI systems shall be evaluated using the quantitative indicators based on fair decision-making capabilities.** At present, specific operations may include:

- Collecting data from trustworthy and legitimate sources to ensure data source trustworthiness.
- Checking the accuracy and completeness of samples, characteristics, and labels in data sets through statistical methods or related toolsets and making timely adjustments based on the check results.
- Updating data sets according to the changes of the real deployment environment to ensure the timeliness and relevance of data sets.
- Building easy-to-use data set formats and interfaces to simplify the process of reading and calling data sets, preventing mal-operation.
- Selecting appropriate quantitative indicators according to specific application scenarios and requirements and considering both individual fairness and group fairness indexes when quantitatively analyzing the capabilities of AI models to make fair decisions.

TABLE 1 COMMON INHERENT BIASES IN DATA SETS

S/N	Data quality	Description
1	Reporting bias	Manually recording of the dataset collection cannot accurately reflect the real and objective situations
2	Automation bias	Results generated by automated software tools are inherently biased
3	Selection bias	Selected samples from the datasets fail to reflect the true distribution of the samples
4	Group attribution bias	People tend to generalize the real situations of individuals to the whole group to which they belong
5	Implicit bias	Assumptions are often made based on models and personal experience that are not necessarily universally applicable

Source: Data compilation

(3) Operation and service stage

In the operation and service stage, enterprises must do well in the explanation of AI systems, monitor the various trustworthiness risks continuously, and actively optimize the systems.

Disclose the technical intents of AI systems to the users. As algorithm explainability is still maturing, the understanding of the technical intents of AI systems may start from the following aspects: establishing appropriate human-machine communication mechanisms; disclosing the functional logical and application requirements of system decision-making; showing the potential risks of wrong decisions. Specifically, when deploying AI systems online, enterprises must establish appropriate human-machine communication mechanisms. For instance, they can establish a function module to inform users if they are interacting with AI systems through an easy-to-understand manner of expression such as text, visual signage, and voice prompts. In practice, users must also be informed of basic information on AI systems, including basic functions, performance, operational requirements, targeted subjects, and the systems' roles in the decision-making process.

Carry out AI risk monitoring continuously. Enterprises must establish user feedback channels to collect authentic user feedback in time and optimize and revise the whole system accordingly. Furthermore, they must monitor the various risks associated with AI systems in practical use, continue to perfect the supervision and compensation mechanisms, determine the liability of AI systems that have caused actual damage and carry out compensation accordingly.

4.2. Industry-level

Apart from the enterprises' efforts, the realization of trustworthy AI also requires the participation and collaboration of various stakeholders. A healthy ecosystem of mutual influence, mutual support, and mutual dependence must be formed. Such an ecosystem should specifically include standardization, assessment & verification, and cooperation & exchange.

The first thing is to build a standard system of trustworthy AI. Policies and laws can only set principles and bottom lines, while trustworthy AI also needs standards to provide specific guidance and constraints from the implementation level. By now, some countries have begun formulating or promulgating principles or laws on AI governance. On this basis, the industry may formulate specific standards and norms in combination with AI technologies, products, or scenarios. For example, the *China National Standard Information Security Technology - Requirements for Security of Face Recognition Data* has begun to solicit public opinions since April 2021. It aims to solve the problems related to facial data collection, such as excessive collection, leakage or loss, and excessive storage and use. Additionally, it has elaborated and refined the facial recognition-related provisions in the *Personal Information Protection Law* draft.

The second is to conduct third-party assessment and verification. Third-party assessment and verification is an effective means to check whether the targeted subjects meet the relevant requirements. Due to its complexity, expert

support from third-party institutions is crucial to AI technology. The assessment and verification must focus on the security, robustness, reproducibility, data protection, traceability, and fairness of the systems based on the characteristics of trustworthy AI. In 2020, China Artificial Intelligence Industry Alliance published the trustworthiness assessment results of the first crop of commercial AI systems, involving 16 AI systems from 11 enterprises, providing an important reference for users' model selection. The EU's latest proposal on AI legislation also proposed measures including AI trustworthiness assessment by authoritative third parties.

The third is to explore a market-based insurance mechanism. Like other information systems, issues associated with AI systems can never be completely resolved, no matter the level of safeguard reached. Therefore, innovations in working procedures are required, to transfer risk by other means. Insurance is an ideal choice. It can shift risks and make up for the users' losses to the greatest extent. Therefore, AI enterprises and insurance institutions should explore an insurance mechanism suitable for AI product use, conduct a quantitative assessment on incident risk, and provide risk compensation jointly, helping to perfect the trustworthy AI ecosystem.

5. Suggestions for trustworthy AI development

Building trustworthy AI systems has become the focus of attention and the direction of efforts from various stakeholders. Adherence to the trustworthy AI methodology will help improve the trustworthiness level of AI in the whole industry and help the public to embrace AI. **It should be noted that trustworthy AI is not finalized. It will continue evolving with the development of AI technologies, ethics, and laws to adapt to the requirements of new advancements.** Such an evolution will also put forward new requirements for all the associated components.

5.1. The process of AI regulation and legislation in China shall be accelerated at the Governmental level

A systematic AI legal regulatory framework should be established. Firstly, the Government should improve the existing laws and regulations to meet developmental requirements. Based on the Cybersecurity Law, Data Security Law, and the Personal Information Protection Law, the issues faced in the process of AI system supervision should be sorted out and the laws and regulations should be updated accordingly. Secondly, the Government should accelerate the process of creating new legislation to respond to new risks actively. The Government should actively explore new AI problems and new AI development trends, rapidly addressing them through the formation of new legislature. Thirdly, the Government should adopt innovative measures to promote the implementation of laws. Measures include the exploration and adoption of pilot, sandbox and other regulatory methods, the development of

intelligent regulatory tools, and the constant improvement of supervisory efficiency and flexibility. Additionally, the Government should continue to oversee the promotion of AI governance through national and international laws, actively participate in multiple bilateral regional cooperation mechanisms, promote the formulation of international rules for AI governance, seek consensus amongst the various parties and bridge differences between parties.

5.2. A comprehensive systematic, and forward-looking layout shall be made at the academic level

Trustworthy AI integration will be an important trend of future research. Current trustworthy AI research focuses on issues of stability, privacy and fairness from a single perspective. Studies have shown that mutually collaborative and restrictive relationships exist between the trustworthy AI requirements of stability, fairness, and explainability. Consideration of only one aspect of trustworthy AI in isolation may lead to conflict with the other requirements. Therefore, it is necessary to build an integrated research framework for trustworthy AI to maintain a dynamically optimal balance among the different characteristic elements.

Research on trustworthy artificial general intelligence (AGI) needs to be laid out in advance. Efforts in AI governance and trustworthy AI mostly focus on weak AI technology and its current applications. General AI and superintelligence receive insufficient attention. However, their development is guaranteed to alter the course of humanity. Therefore, it is necessary to adopt a forward-looking strategy and explore the development path of general AI by developing cutting-edge technologies, including super-deep learning and quantum machine learning. Furthermore, trustworthiness-related research must accompany the exploration of strong AI.

5.3. Enterprises' practices must adapt to the business development and achieve agile trustworthiness

Enterprises must pay attention to the agile iteration of trustworthy AI while expanding the applications of AI technology. With the extensive integration of AI technology with different industries, their practical usage will put higher trustworthiness requirements on enterprises. On the one hand, enterprises must develop trustworthy AI testing and monitoring tools to adapt to the business development needs, upgrading and iterating them based on the characteristics of industry applications. On the other hand, they must actively establish a rapport with regulatory authorities and cooperate with regulatory measures including digital sandbox, safe haven, application trials, and the compliance of standards, to build an agile trustworthiness mechanism that can coordinate internal and external resources.

5.4. A communication and cooperation platform shall be set up to create a trustworthy ecosystem at the industry organization level

The industry organizationis encouraged to build a dedicated communication platform for trustworthy AI and call on all parties in the industry to build a trustworthy AI ecosystem jointly. Trustworthy AI is a complex and systematic project, which requires the participation of multiple stakeholders. The industry organization should fully utilize their advantage to: extensively learn and incorporate exemplary practical experience from various parties to prepare operational guidelines for trustworthy AI; establish and improve trustworthy AI standards based on the aspects of R & D, management, technical support and product application; accelerate the research and development of AI evaluation and monitoring capabilities, and continuously promote the implementation of trustworthy AI in industrial circles through various means including testing, tracking and monitoring.

References

[1] EUROPEAN COMMISSION. *WHITE PAPER On Artificial Intelligence-A European approach to excellence and trust*[R/OL]. (2020-02-19) https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[2] Jobin A., et al. *The global landscape of AI ethics guidelines*[J]. Nature Machine Intelligence, 2019, 1(2).

[3] 中国人工智能产业发展联盟. *人工智能行业自律公约*[R/OL].(2019-0808) http://aiiaorg.cn/uploadfile/2019/0808/20190808053719487.pdf

[4] 中国人工智能产业发展联盟. *可信 AI 操作指引*[R/OL].(2020-0923) http://aiiaorg.cn/uploadfile/2020/0923/20200923064427421.pdf

[5] 张钹等. *迈向第三代人工智能*[J]. 中国科学:信息科学, 2020, v.50(09):7-28.

[6] 何积丰. *安全可信人工智能*[J]. 信息安全与通信保密, 2019(10):4-8.

[7] Liu T., et al. *Algorithm-dependent generalization bounds for multi-task learning*[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 39, pages 227-241, 2016.

[8] He F., et al. *Control batch size and learning rate to generalize well: Theoretical and empirical evidence*[C]. In Advances in Neural Information Processing Systems, pages 1141-1150, 2019.

[9] Tu Z., et al. *Theoretical analysis of adversarial learning: A minimax approach*[C]. In Advances in Neural Information Processing Systems, pages 12280–12290, 2019.

[10] Dwork C., et al. *The algorithmic foundations of differential privacy*[J]. Foundations and Trends in Theoretical Computer Science, volume 9, pages 211–407, 2014.

[11] Zhu, L., et al. *Deep leakage from gradients*[C]. In Advances in Neural Information Processing Systems, 2019.

[12] He F., et al. *Tighter generalization bounds for iterative differentially private learning algorithms*[J]. arXiv preprint arXiv:2007.09371, 2020.

[13] Dwork C., et al. *Fairness through awareness*[C]. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226, 2012.

[14] Ribeiro M. T., et al. *"Why should i trust you?" Explaining the predictions of any classifier*[C]. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135-1144, 2016.

[15] Ribeiro M. T., et al. *Anchors: High-precision model-agnostic explanations*[C]. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[16] Calders, T., et al. *Building classifiers with independency constraints*[C]. IEEE International Conference on Data Mining Workshops. pages 13-18, 2009.

[17] Fang, M., et al. *Poisoning attacks to graph-based recommender systems*[C]. In Proceedings of the 34th Annual Computer Security Applications Conference, pages 381-392, 2018.

[18] Eykholt, K., et al. *Robust physical-world attacks on deep learning visual classification*[C]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1625-1634, 2018.

[19] McMahan, B., et al. *Communication-efficient learning of deep networks from decentralized data*[C]. In Artificial Intelligence and Statistics, pages 1273-1282, 2017.

[20] Hardt, M., et al. *Equality of opportunity in supervised learning*[C]. In Advances in Neural Information Processing Systems, pages 3323-3331, 2016.

China Academy of Information and Communications Technology

Address: No.52, Hua Yuan Bei Road, Haidian District, Beijing, China

Postcode: 100191

Tel: 010-62309514

Fax: 010-62304980

Website: www.caict.ac.cn